Review of "A North Sea in situ evaluation of the FitchWind Farm Parametrization within the Mellor–Yamada–Nakanishi–Niino and 3D Planetary Boundary Layer schemes" by Agarwal et al.

The manuscript compares a new planetary boundary layer scheme to the standard one used for wind farm modeling in WRF. It evaluates the accuracy of the PBL schemes with offshore measurements of two independent sources. A thorough analysis is carried out with respect to model setup and statical metrics.

This work addresses an understudied, yet extremely relevant aspect of mesoscale wind farm modeling. However, several aspects of the manuscript need major improvement before acceptance.

Specific comments

- The authors should significantly narrow the scope of the work to make the analysis more comprehensible and easier to read. Since the main focus is to compare 3DPBL with MYNN, all results not directly supporting this comparison should be removed from the analysis. Specifically:
    - Varying the TKE factor is a relevant analysis, but clouds the clarity of comparison between PBL models. Simply choose one TKE factor (suggestion: default 0.25) and present only these results in figures 5-15. Then add a final section at the end of the results that focuses on the effect of the TKE factor, illustrated by one figure.
    - The appendix should serve the manuscript by providing extra explanations or results that didn't fit to the story, whereas currently the appendix feels more like a 'result dump'. It should be revised what results are interesting enough to add to the appendix.
- The conditions on which the evaluation takes place should be more clearly described. Currently, the combination of four statistical metrics, three data sets (or more if you separate between FINO heights) and six WRF runs (2 PBLs times 3 TKE factors) is extremely overwhelming when trying to find the main result. Especially since the results are not very consistent, this results in a very chaotic discussion and conclusion, see for instance the abstract and l. 316-324. You claim that the 3DPBL is better than MYNN in the rotor area (l. 344-346), which is not that clear from the results shown before.
- The differences between the two PBL schemes should be related to the physics that are modeled in these schemes. In the discussion of the results, it is currently mostly stated that there is a difference, but an interpretation on why this makes sense is missing. Besides, a paper comparing different PBL schemes should definitely include a section in the methodology that briefly described the fundamental principles of these schemes, possibly supported by the most relevant equations. This can then be referred to explain the results. Some examples:
    - L. 235-236: "MYNN simulations … the surface."
    - L. 249-250: "Both the … simulation (Fig. 6a,c)."
    - L. 261-263: "MYNN average … wind speeds."
    - L. 294-295: "TKE in … MYNN simulations."

- The manuscript draws some very thin conclusions that are not (fully) supported by the results or explained in the accompanying text. Some examples:
    - l. 268-270: You say that the wind speed in the wakes in MYNN are stronger because the higher inflow wind speed. However, the difference is very large (up to 1 m/s). I suspect there is more to it than just the difference in inflow wind speed, especially since Fig. 7a shows that the difference in WS offshore are not that large. To proof your claim, I'd recommend comparing the wakes under the same inflow wind speed. This can be done by either running WRF over a longer period of time and binning the results, or designated idealized WRF runs. This would help greatly in explaining the remainder of your results: is the difference between the PBL schemes due to a change in estimated background wind speed, or because of wake dissipation?
    - L. 344-346: You claim that your results support that 3DPBL is better than MYNN in the rotor area, although the results of FINO1 show that this is highly dependent on the statistical metric used.
    - L. 399-406: You claim that a TKE factor of 1 is better for MYNN in your study. However, to claim this you need to be absolutely sure that the background (without turbines) wind speed and TKE are correct. It could be possible underestimates WS and TKE, which are now corrected by adding more turbulence. This doesn't mean that this is physically correct. Besides, Fig. 13-15 clearly shows that for both PBL schemes the trend is the same, there is just a systematic offset. As the TKE factor represent a physical process, it is wrong to conclude that one TKE factor is better for MYNN and other for 3DPBL.
- A discussion on the uncertainty of your results is completely absent. A critical evaluation of your results to the uncertainty in the background wind speed is essential for the interpretation of the results. You show that the background wind speed is about 1 m/s higher in MYNN, but then only evaluate on the waked wind speed. Is 3DPLB then really better, or is it just cancelling out errors?
  Additionally, please address the effect of model grid resolution and generalizability to other sites and longer time series.

Technical corrections

- l. 45-46: "To date" is followed by a reference to a paper from 2014. Update this
- l. 53-65: This section states what other papers have carried out but does not discuss their findings.
- l. 72: somewhere here the 3D TKE scheme should also be mentioned, and compared (qualitatively) to 3D PBL.
- L. 82-87: It is unclear how this section contributes to the storyline in the introduction
- Fig. 2: ensure that the axes are equal, which should result in square grid boxes.
- L. 130-131: mention exactly what WRF version
- L. 143: "[]" should be "[-]"
- L. 153: detail like file name not needed

- Fig. 4: added benefit of showing this figure unclear
- L. 187-188: Do these studies use the same grid size as you do? Or is the 2km they recommend based on their grid resolution?
- L. 217-219: provide equations for EMD like done for all other metrics
- L. 231-232: add Introduction to help the reader prepare for what's coming
- L. 233: header not really appropriate for this section
- L. 237: wind direction in plot is shown to rotate from southwest to west, so it veers as is expected in SBL
- Fig. 5: why are no measurement added here? Or alternatively, why not use the NWF runs for a cleaner analysis of the profiles?
- L. 244: Add PBLH as reference
- Fig. 6: why are there only contour plots for TKE, and not for WS? Suggest to either add WS contour plots, or just add TKE profile to Fig. 5.
- Sect. 3.2.1: Why is there not TKE/TI comparison for FINO1?
- L. 281-282: increasing TKE factor increases wind speed
- L. 287-288: what should be noted is that the wind speeds a lower altitudes are also lower, resulting in lower biases. Alternatively look at relative biases
- Fig. 8: unclear what the mean values are
- Fig. 11: relate findings to FINO results at multiple heights: can these results be extrapolated to arrive at the results shown here? Additionally, 11c shows the opposite to FINO, please explain why.
- L. 307: The issue of relative sensor alignment is very relevant and deserves more attention than this single sentence. Mention this in the methods section when introducing the aircraft measurements and include it in the discussion of uncertainties.
- L. 355: Fig. 14a is aircraft, but text says FINO1.
- L. 363: Provide references for "recent scientific discussion"
- L. 433-447: This outlook does not belong to the conclusion section.
- L. 461-462: If you claim this, show that the 3DPLB scheme indeed extracts more momentum for higher altitudes.
- L. 509-523: Larsen & Fishereit is mentioned 7 times in just a few sentences. Rewrite this. This happens all over the manuscript, please review that you don't insert unnecessary references.
- Appendix A3: You show here that the results shown in the main body of the manuscript hold for just the 2h aircraft data is available. If this is true, why show the 12h analysis at all? Explain in the main body of the text why you argue that the 12h analysis is better. Additionally, FINO1 allows for an analysis at a much longer time scale. Justify why you restrict your analysis to just 12 hours.